Delay Guaranteed Joint User Association and Channel Allocation for Fog Radio Access Networks

Minglei You, Gan Zheng *Fellow, IEEE*, Tianrui Chen, Hongjian Sun, *Senior Member, IEEE* and Kwang-Cheng Chen, *Fellow, IEEE*

***Abstract*—In the Fog Radio Access Networks (F-RANs), the local storage and computing capability of Fog Access Points (FAPs) provide new communication resources to address the latency and computing constraints for delay-sensitive applications. To achieve the ultra-low latency, a novel joint user association and channel allocation scheme is proposed in this paper, where the FAPs are clustered from a user-centric perspective. The delay performance is improved regarding both the control signaling procedure and the data transmission procedure. Specifically, the multiple access interference (MAI) between users is analyzed, where the closed-form expression for the effective rate of a typical user with multiple FAP connections and arbitrary interfering users is obtained. With the consideration of MAI, the proposed distributed joint user association and channel allocation algorithm provides a guaranteed delay violation probability. Moreover, the distributed algorithm can be conducted on individual FAPs, whose calculation is simplified by look-up tables. Simulation results show that the proposed algorithm is capable of providing statistical delay performance guarantee including both average delay and delay bound violation probability, which demonstrates its superiority in supporting delay-sensitive applications in F-RANs.**

***Index Terms*—User Association, Channel Allocation, Effective Capacity, Fog Radio Access Networks, 5G, 6G.**

I. Introduction

The advancement of smart devices, such as smartphones, high-end wearables and connected vehicles, are boosting the demand for massive wireless connections of ultra-low latency for real-time services [1] [2]. It is well known that the Cloud Radio Access Networks (C-RANs) have been proposed to further extend throughput and coverage in the emerging fifth-generation (5G) wireless communication systems. In the C-RANs, the traditional base station (BS) is decoupled into distributed remote radio heads (RRHs) for radio services and baseband units (BBUs) for baseband signal processing [3]. Even though the centralized BBU can be very powerful in signal processing, the fronthaul link's capacity between RRHs and the BBU becomes the bottleneck as the network scaling up [4]. Consequently, Heterogeneous C-RANs (HC-RANs) are proposed in [5], where high power nodes provide seamless coverage and execute the functions of the control plane, while RRHs are providing high-speed data transmissions in the user plane. However, both C-RANs and HC-RANs are facing challenges in supporting delay-sensitive tasks due to the possible long network delay in transmissions to the centralized BBUs [6]. Situations could get worse with the increasing demand for uplink data transmissions, where Cisco estimated that by 2021 the data generated at the edge will boost to 850 ZettaBytes (ZB), while the global data center traffic will reach 20.6 ZB [7]. This trend demands the evolution of existing centralized networks to a more distributed architecture, and correspondingly the network design and algorithms are needed to support the distributed computing and storage features.

To address the above challenges, the Fog Radio Access Networks (F-RANs) are evolved based on HC-RANs, where the Fog Access Points (FAPs) are equipped with storage and computing capacities [3]. The FAPs provide new communication resources in supporting delay-sensitive applications, where the FAPs can be used to assist the BBUs for computing [8] or clustered as mini clouds to serve users [6]. Moreover, with the advantage of distributed deployment, multiple FAPs can be dynamically clustered to serve individual users, which is referred to as the user-centric method. The user-centric method provides new approaches to allocate resources in the networks on demand. In [9], a joint user association and scheduling algorithm is proposed to optimize the system throughput, which decouples the load balancing problem via the iterative cooperation among multiple distributed base stations. A user-centric dynamic self-organizing method is proposed in [10], where the user clustering problem is optimized in an iterative way with the consideration of load balancing. The delay-aware joint caching and user association optimization is studied in [11], where the effective capacity theory is used to characterize the end-to-end latency. However, the distributed user association algorithms lack centralized coordination and global information, which are commonly relying on iterative procedures to make the final decision as exploited in the aforementioned works. In this paper, to resolve the long delay due to the iterative procedure, we propose a distributed algorithm by decoupling the original problem to local computing with the pre-cached regional user service status.

One key 5G application scenario is the support of real-time applications, such as video streaming, cloud gaming and augmented/virtual reality. Therefore the resource allocation via Coordinated Multi-Point (CoMP) transmission in support of ultra-low latency has attracted attentions in the studies

Minglei You and Hongjian Sun are with Department of Engineering, Durham University, UK, DH1 3LE (e-mail:{minglei.you, hongjian.sun}@durham.ac.uk). Gan Zheng and Tianrui Chen are with Wolfson School of Mechanical, Electrical and Manufacturing Engineering, Loughborough University, UK, LE11 3TU (e-mail:{g.zheng, t.chen}@lboro.ac.uk). Kwang-Cheng Chen is with the Department of Electrical Engineering, University of South Flordia, USA, FL 33620. (e-mail: kwangcheng@usf.edu).This work was supported in part by the UK EPSRC under grant number EP/N007840/1 and Leverhulme Trust Research Project Grant under grant number RPG-2017-129. K.-C. Chen was supported in part by a grant from Cyber Florida.

of C-RANs, HC-RANs and F-RANs. The downlink resource allocation and user association is optimized for the packet delay performance using the Alternating Direction Method of Multipliers [12]. The delay-aware uplink fronthaul allocation problem is studied in C-RANs [13], where low-complexity algorithms are proposed with asymptotic approximations. The downlink CoMP system is studied in [14], where the interference alignment and neutralization techniques are used to improve the spectral efficiency for the cell-edge users. The non-orthogonal multiple access (NOMA) based techniques is proposed in [15], which is designed to improve the system throughput in the uplink CoMP systems. In [16], the user association is optimized based on the queue status in HC-RANs. With the feasible local computing capacity in the FAPs, it enables the shift of traditional centralized computing tasks to the network edges near the users [17]. Especially for the delay-sensitive applications, the shift of computing and caching to the logical edge of the network shows promising solutions in reducing the latency. In [3], a joint distributed computing scheme is proposed to cluster the FAPs, which optimizes the overall latency by minimizing transmission delays and computing delays. A latency-driven fog cooperation approach is proposed in [6], which dynamically selects the FAPs to achieve a trade-off between the computing time and communication time. The cooperation between FAPs and BBUs is optimized for workload allocation in [8], which provides a trade-off between the power consumption and transmission delay. However, as the FAPs are dynamically clustered to serve users in the user-centric based F-RANs, there will be coverage overlap which is different from traditional cellular infrastructure. Therefore the decisions on the new users' association and channel allocation will also influence the delay performance of existing users. This requires an effective resource coordination between distributed FAPs, where the statistical delay provision for all users in the network should be addressed.

To address the aforementioned challenges, a joint user association and channel allocation scheme is proposed for the uplink of F-RANs, which provides a distributed solution for the delay-sensitive applications. The main contributions are summarized as follows:

- A novel joint user association and channel allocation scheme is proposed, which provides a statistical delay guarantee for the delay-sensitive applications in the F-RANs. The FAPs are clustered as a virtual cell from a user-centric aspect, while existing users' delay performance is protected during the new user's association and channel allocation. Specifically, the influence between the new user and existing users are considered in a mutual way. The new user's association and channel allocation decision is based on the service status of the existing FAPs and users, while its influence on the existing FAPs and users is also evaluated. The scheme improves not only the delay performance during the data transmission phase, but also the association procedure regarding the control signaling procedure.
- The effective rate analysis is performed on typical users with multiple FAP associations and arbitrary interfering users, while a closed-form expression is obtained in the Rayleigh channel condition. This analytical result can be exploited in general multi-user scenarios, where each user is served by multiple FAPs and each FAP is serving multiple users. Meantime, the closed-form expression can be extended to other fading conditions such as Nakagami-m fading channels for wider application scenarios.
- To the authors' best knowledge, this work is the first to use the effective rate as a tool to allocate resources for statistical performances in multi-user wireless communication systems. This is a novel effort in the effective rate as well as resource allocation related studies, which is different from the existing works on effective rate calculations, system performance evaluations or resource allocations for effective rate maximization.
- The distributed computation and storage constraint aspects are addressed during the design of the algorithm. Specifically, each FAP is capable to make regional decisions based on the pre-cached regional service status and the instantaneous local user information. This feature not only enables the FAPs to scale up for larger networks, but also avoids the latency due to the cooperation between FAPs during the user association and channel allocation. Moreover, a lookup table method is proposed to transform the algorithms into a computationally effective format, which addresses the limited computing capability issue of the FAPs.

The remainder of this paper is structured as follows. In Section II, the system model is introduced and the statistical Quality of Service (QoS) analyses of a typical user with multiple FAPs and arbitrary interfering users is performed. Based on these results, the joint user association and channel allocation scheme is proposed in Section III, and numerical evaluation results are presented in Section IV. The conclusions are drawn in Section V with extended discussions.

Throughout the paper, following notations are used. Bold letters donate vectors. $\mathbf{1}_N$ denotes all-one vector of N elements. $f_\gamma(z)$ denotes the probability density function (PDF) of γ, while $\phi_\gamma(s)$ denotes the moment generating function (MGF) of γ. $\mathbb{E}_g\{\cdot\}$ denotes the expectation operator of the random variable g. $\Gamma(z)$ is the gamma function [18, (5.2.1)]. $H_{p,q}^{m,n}[\cdot]$ [19] denotes the uni-variate or multi-variate H function depending on the variables as detailed in Appendix A.

II. System Model and Statistical QoS Analyse in F-RANs

A. System Model

In this paper, an uplink interference-limited F-RAN scenario is considered as illustrated in Fig. 1. The distributed single-antenna FAPs are equipped with local computing capability and storage capacity. The links among FAPs as well as BBUs are fulfilled by high-speed connections, e.g., Gigabit Ethernet. It is assumed that there exists a networking mechanism using High Power APs to serve all users within the network to provide seamless coverage and service. The received signals are combined at one selected FAP via the maximum-ratio combining (MRC) method, which offloads the data transmission

between FAPs and BBUs to the links between regional FAPs. In this paper, the MRC method is exploited because it can be implemented in a distributed manner with low implementation complexity [20]. As latency is the main focus in this paper, the open-loop power control is applied due to its potential latency advantage comparing to the closed-loop power control.

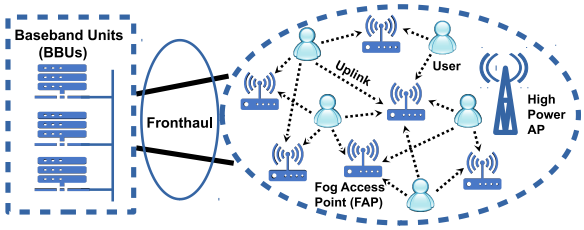

Fig. 1. An illustration of the considered F-RAN network, where the FAPs are clustered via the user-centric method.

In the user-centric clustering method, each user can access multiple FAPs, while each FAP can simultaneously serve different regional clusters. Therefore users on the same channel will introduce multiple access interference (MAI) to each other. Let $\{\text{AP}_l\}$ denote FAPs within a distance ρ_{FAP} of the typical user U_0, where $l = 1, \ldots, L$. It is assumed there are J channels and the orthogonal frequency division multiple access (OFDMA) is used. For each AP_l, there are K_l^j users on the channel j, whose distance to AP_l is denoted by $r_{lk_l^j}$. Since each user is allocated with one channel and the interference is from the co-channel users, the superscript j is omitted for notation simplicity where there is no ambiguity.

In the following, we will study the performance of a typical user U_0, whose distance to AP_l is r_{l0}. The received power S_l at AP_l from U_0 can be given as follows:

$$S_l = P n_{l0} g_{l0}, \tag{1}$$

where P is the user's average transmit power, g_{l0} is the fast fading parameter, and $n_{l0} = r_{l0}^{-\alpha}$ is the large scale fading parameter, while α is the path loss exponent. Similarly, the interference at AP_l introduced by U_{k_l} can be given as follows:

$$I_{lk_l} = P n_{lk_l} g_{lk_l}, \tag{2}$$

where $n_{lk_l} = r_{lk_l}^{-\alpha}$. The co-channel users could interfere multiple APs serving U_0, where the interference powers are counted at each AP accordingly. Thus the aggregated interference at AP_l can be given by

$$I_l = \sum_{k_l=1}^{K_l} I_{lk_l}. \tag{3}$$

Therefore for U_0 at AP_l, the signal-to-interference ratio (SIR) γ_l can be given by

$$\gamma_l = \frac{S_l}{I_l}, \tag{4}$$

where the noise effect is ignored due to the strong interference from other users. Since P is the common factor that can be canceled out in the SIR definition of (4), it is assumed that $P = 1$ for notational simplicity in the rest of the paper. Moreover, Rayleigh fading with unit variance [16] is used to describe the fast fading effect experienced by the communications between FAPs and users. The PDF of S_l can be given as follows:

$$f_{S_l}(z) = \frac{1}{n_{l0}} e^{-\frac{z}{n_{l0}}}. \tag{5}$$

Similarly, the PDF of I_{lk_l} can be given by

$$f_{I_{lk_l}}(z) = \frac{1}{n_{lk_l}} e^{-\frac{z}{n_{lk_l}}}. \tag{6}$$

As the FAPs are spatially separated, it is assumed that I_{lk_l} as well as S_l are mutually independent. Therefore the MGF of γ_l in (4) can be given by the following proposition.

Proposition 1: The MGF of the SIR γ_l of the typical user U_0 with multiple FAP connections can be given in a closed form as follows:

$$\phi_{\gamma_l}(s) = \sum_{k_l=1}^{K_l} w_{lk_l} H_{1,2}^{2,1}\left[\frac{n_{l0}}{n_{lk_l}} s \middle| \begin{matrix}(0,1)\\(0,1),(1,1)\end{matrix}\right], \tag{7}$$

where the uni-variate H function $H_{p,q}^{m,n}[\cdot]$ [1] is defined in the Appendix A and w_{lk_l} is given by

$$w_{lk_l} = \prod_{i=1, i\neq k_l}^{K_l} \frac{1}{1 - \frac{n_{li}}{n_{lk_l}}}. \tag{8}$$

Proof: See Appendix B for details. ■

Here the MRC method is used to combine the signals from different FAPs associated with U_0, therefore the MGF $\phi_{\gamma_{\text{end}}}(s)$ of the aggregated SIR $\gamma_{\text{end}} = \sum_{l=1}^{L} \gamma_l$ can be given as follows:

$$\phi_{\gamma_{\text{end}}}(s) = \prod_{l=1}^{L} \phi_{\gamma_l}(s). \tag{9}$$

B. Effective Rate Analysis of Typical Users in F-RANs

In this part, the effective rate is used to quantify the statistical delay performance, which has been exploited in a wide range of resource allocation problems, e.g., multiple-input multiple-output (MIMO) systems [21], multi-hop systems [22], and power systems [23], where a comprehensive survey can be found in [24]. For a fading channel, the effective rate is defined as the maximum constant rate it can support under statistical delay constraints [25] as follows [26]:

$$R(\theta, \boldsymbol{n}) = -\frac{1}{\theta BT} \ln \mathbb{E}_{\boldsymbol{g}}\{e^{-\theta T C(\boldsymbol{\gamma})}\}, \tag{10}$$

where $C(\boldsymbol{\gamma})$ represents the system service rate during a single time block with duration T and bandwidth B, while θ is the QoS exponent. Here $\boldsymbol{\gamma} = [\gamma_1, \ldots, \gamma_L]$, $\boldsymbol{g} = [g_{10}, \ldots, g_{LK_L}]$ and $\boldsymbol{n} = [n_{10}, \ldots, n_{L0}]$ are used for simplicity. The QoS exponent θ is given by [26]

$$\theta = -\lim_{z\to\infty} \frac{\ln \Pr\{Q > z\}}{z}, \tag{11}$$

where Q is the equilibrium queue-length of the buffer. The probability $\epsilon(\theta, \boldsymbol{n})$ of the delay D larger than the delay bound $D_{\max}$ can be estimated by [21]

$$\epsilon(\theta, \boldsymbol{n}) \triangleq \Pr\{D \geq D_{\max}\} = e^{-\theta B R(\theta, \boldsymbol{n}) D_{\max}}, \tag{12}$$

[1] Note that since multi-variate H function reduces to uni-variate H function when there is only one random variable, the presentation is unified to $H_{p,q}^{m,n}[\cdot]$ and can be distinguished by the number of involved random variables.

where θ is the value required to support a constant data arrival rate of λ as given below [22]

$$R(\theta, \boldsymbol{n}) = \frac{\lambda}{B} \triangleq R_{\min}, \tag{13}$$

where the constant $R_{\min}$ is used for simplicity. Note that the study with the constant arrival rate traffic can also serve as a worst-case bound for other practical traffic types such as the sporadic or burst traffic model, which can be proved by comparing the QoS exponents and using Lemma 1 from [27]. A relation between ϵ and θ can then be obtained by substituting (13) into (12) as follows:

$$\theta = \frac{-\ln \epsilon}{\lambda D_{\max}}. \tag{14}$$

Note that θ is a parameter determined by both the wireless transmissions (characterized by the effective rate function $R(\theta, \boldsymbol{n})$) and the data traffic as given in (13). The resource allocations for wireless communication systems, e.g., the user association and channel allocation, will affect the effective rate function $R(\theta, \boldsymbol{n})$, which leads to the change of θ via (13) and consequently impact on the statistical delay violation probability ϵ in (12). Therefore θ is named as QoS exponent, which serves as a metric in the study of statistical delay guarantees. With (10) – (14), some important properties regarding $R(\theta, \boldsymbol{n})$ and $\epsilon(\theta, \boldsymbol{n})$ can be derived in the following lemma.

Lemma 1: The effective rate $R(\theta, \boldsymbol{n})$ is a monotonically decreasing and concave function of the QoS exponent θ, while the statistical delay violation probability $\epsilon(\theta, \boldsymbol{n})$ is a monotonically decreasing and convex function of the QoS exponent θ, i.e., the following properties hold:

$$\frac{\partial R}{\partial \theta} \leq 0, \ \frac{\partial^2 R}{\partial \theta^2} \leq 0, \ \frac{\partial \epsilon}{\partial \theta} \leq 0, \ \frac{\partial^2 \epsilon}{\partial \theta^2} \geq 0, \ \forall \theta > 0. \tag{15}$$

Proof: The proof follows Lemma 1 in [23] and Proposition 2 in [21]. ∎

From Lemma 1 and (10) – (14), it can be inferred that a smaller θ corresponds to a more stringent QoS performance. Meantime, when $\theta \to 0$, the effective rate approaches the Shannon channel rate. With Proposition 1, the effective rate of the typical user can be given as follows.

Proposition 2: The effective rate of the typical user with multiple FAP associations and arbitrary interfering users can be given in a closed form as follows:

$$R(\theta, \boldsymbol{n}) = -\frac{1}{A} \log_2 \frac{1}{\Gamma(A)} \sum_{k_1,\ldots,k_L} H_{1,0}^{0,1}\left[\begin{matrix}(1-A, \mathbf{1}_L):\\ -\end{matrix} \left(\frac{1}{L}, \mathbf{O}_{k_l}, \mathbf{P}_{k_l}\right)_{1,L}\right] \tag{16}$$

where $A = \frac{\theta T B}{\ln 2}$. $H_{p,q}^{m,n}[\cdot]$ [19] is the multivariate *H* function detailed in Appendix A and the parameter set $\mathbf{O}_{k_l} = [2, 1, 1, 2]$ and $\mathbf{P}_{k_l} = [w_{lk_l}, \frac{n_{l0}}{n_{lk_l}}, 0, (0, 1), 1, (1, 1)]$. The operation of $\sum_{k_1,\ldots,k_L}$ means the summation over all possible combinations of the listed elements in $k_1, \ldots, k_L$.

Proof: By substituting (7) into (9) and swapping the order of summation and production, the MGF of the typical user can be obtained as

$$\phi_0(s) = \sum_{k_1,\ldots,k_L} \prod_{l=1}^{L} w_{lk_l} H_{1,2}^{2,1}\left[\frac{n_{l0}}{n_{lk_l}} s \middle| \begin{matrix}(0,1)\\ (0,1),(1,1)\end{matrix}\right]. \tag{17}$$

By applying Theorem 1 in [28] and taking short-hand representation of multi-variable *H* function as detailed in Appendix A, (16) can be obtained. ∎

The result (16) is novel, which is in a closed form and general for users with multiple FAPs connections and arbitrary interfering users. This highly structured representation via *H* function is beneficial for calculation, where the parameters are well grouped according to interfering users. Moreover, although the discussion is based on Rayleigh fading channel, the results of Proposition 1 and Proposition 2 can be extended to other fading scenarios that can be described by Gamma random variables [29], e.g. Nakagami-m fading channels that can characterize a wider range of fading channels, including one-sided Gaussian, Rician and log-normal fading [30].

III. Joint User Association and Channel Allocation Scheme

The association of a new user to multiple FAPs is a complex problem. On the one hand, the association and channel allocation should provide a delay guaranteed performance to this new user. On the other hand, the new user will add new interference to all existing users on the same channel, whose delay performance will be affected. To provide a universal delay guarantee service to all users, including both new users and all existing users, a novel joint user association and channel allocation scheme is proposed in this section.

It is assumed that for all users, the statistical delay performance requirements, i.e., the delay bound $D_{\max}$ and the delay bound violation probability $\epsilon_{\max}$, are the same. For the typical user U_0, its statistical delay guarantee performance can be characterized as follows:

$$\epsilon(\theta_0^*, \boldsymbol{n}) \leq \epsilon_{\max}, \tag{18}$$

where θ_0^* is calculated via (13), i.e., $R(\theta_0^*, \boldsymbol{n}) = R_{\min}$. Unfortunately, there are no general closed-form solutions to obtain θ_0^* from this equation. Traditionally, this is solved by numerical methods such as search algorithms. However, a brute force calculation of θ_0^* is not efficient and may result in extra processing time. In the following, we will transform the statistical delay guarantee in (18) to a more tractable form.

A. Problem Formulation

According to Lemma 1, ϵ is monotonically decreasing with respect to θ. Therefore for the typical user U_0, its delay violation probability constraint (18) is equivalent to the following constraint on the QoS exponent θ_0^*:

$$\theta_0^* \geq \theta_{\min}, \tag{19}$$

where $\theta_{\min}$ is a constant given as follows:

$$\theta_{\min} = \frac{-\ln \epsilon_{\max}}{\lambda D_{\max}}. \tag{20}$$

Note that although the actual QoS exponent varies with different users, the lower bound of QoS exponent $\theta_{\min}$ is a constant determined by each users' QoS requirements. Further we see the effective rate function $R(\theta, \boldsymbol{n})$ is monotonically decreasing with respect to θ as given in Lemma 1. Therefore for the typical user U_0, (19) has an equivalent form as follows:

$$R(\theta_{\min}, \boldsymbol{n}) \geq R(\theta_0^*, \boldsymbol{n}) \triangleq R_{\min}, \tag{21}$$

where in (21), the equation part is due to the fact that θ_0^* is the value satisfying (13). In this way, for every user association and channel allocation decision, we don't need the exact θ_0^* to calculate its delay performance $\epsilon(\theta_0^*, \boldsymbol{n})$ and check whether it can satisfy (18). Instead, we can calculate the effective rate with given user association and channel allocation decisions at the fixed point $\theta_{\min}$, and the delay performance is guaranteed if (21) is satisfied, where the effective rate function under different QoS exponents has been given in closed form in (16).

In this way, the effective rate function $R(\theta, \boldsymbol{n})$ is used as a performance indicator of the user association and channel allocation decisions. For the typical user U_0, it is desirable to have $\epsilon(\theta_0^*, \boldsymbol{n}) - \epsilon_{\max}$ as small as possible, which is equivalent to a statistical delay guarantee as good as possible. Using Lemma 1, $\epsilon(\theta_0^*, \boldsymbol{n}) - \epsilon_{\max}$ can be one-to-one mapped to the effective rate margin $R(\theta_{\min}, \boldsymbol{n}) - R(\theta_0^*, \boldsymbol{n})$. Specifically, if the new user is allocated with channel j and the associated FAP cluster is $\mathbb{A}_j$, then the new user's delay performance added to the system is quantified by the effective rate margin $R(\theta_{\min}, \boldsymbol{n}_{\mathbb{A}_j}) - R(\theta_{\mathbb{A}_j}, \boldsymbol{n}_{\mathbb{A}_j})$, where $\theta_{\mathbb{A}_j}$ is the new user's QoS exponent. Meantime, the influence on existing users' delay performance is quantified by the change of their effective rate margin before and after the user association and channel allocation, respectively. As an example, consider an existing user U_i within all interfering users set $\mathbb{U}_j$ on channel j, whose large scale fading parameter set is $\mathbf{n}_{U_i}$ and its effective rate margin is $R(\theta_{\min}, \boldsymbol{n}_{U_i}) - R(\theta_{U_i}, \boldsymbol{n}_{U_i})$. After the user association and channel allocation, its QoS exponent is changed from θ_{U_i} to θ'_{U_i}, its large scale fading parameter set is changed from $\boldsymbol{n}_{U_i}$ to $\boldsymbol{n}'_{U_i}$, while its effective rate function is changed from $R(\theta, \boldsymbol{n}_{U_i})$ to $R(\theta, \boldsymbol{n}'_{U_i})$. Then this change of effective rate margin can be calculated as $[R(\theta_{\min}, \boldsymbol{n}'_{U_i}) - R(\theta'_{U_i}, \boldsymbol{n}'_{U_i})] - [R(\theta_{\min}, \boldsymbol{n}_{U_i}) - R(\theta_{U_i}, \boldsymbol{n}_{U_i})]$.

In this paper, the objective of the joint user association and channel allocation problem is to maximize the system-wise statistical delay performance gain due to potential decisions, which is formulated as the summation of the new users' effective rate margin and the change of existing users' effective rate margin as follows:

$$\max_{j, \mathbb{A}_j} R(\theta_{\min}, \boldsymbol{n}_{\mathbb{A}_j}) - R(\theta_{\mathbb{A}_j}, \boldsymbol{n}_{\mathbb{A}_j}) + \sum_{U_i \in \mathbb{U}_j} \Big\{ \big[R(\theta_{\min}, \boldsymbol{n}'_{U_i}) - R(\theta'_{U_i}, \boldsymbol{n}'_{U_i})\big] - \big[R(\theta_{\min}, \boldsymbol{n}_{U_i}) - R(\theta_{U_i}, \boldsymbol{n}_{U_i})\big] \Big\} \tag{22a}$$

$$\text{s.t.} \quad |\mathbb{A}_j| \leq L_{\max}, \tag{22b}$$

$$R(\theta_{\min}, \boldsymbol{n}_{\mathbb{A}_j}) \geq R_{\min}, \tag{22c}$$

$$R(\theta_{\min}, \boldsymbol{n}'_{U_i}) \geq R_{\min}, \tag{22d}$$

where $\mathbb{U}_j = \{U_1, \ldots, U_{K_j}\}$, $j = 1, \ldots, J$ and U_i is the ith interfering user on channel j, $i = 1, \ldots, K_j$. The constraint (22b) ensures that the number of connected FAPs L doesn't exceed the maximum allowed FAP connection number $L_{\max}$ for the users. Meantime, (22c) and (22d) ensure the delay bound violation probability in (18), which is due to the analysis in (18) – (21). The optimization of the objective function in (22a) depends on two parameters, namely the allocated channel j and the associated FAP cluster $\mathbb{A}_j$. On different channels, the users will encounter different co-channel users, whose interference will be reflected in the received signals of the associated FAP cluster $\mathbb{A}_j$. Therefore it requires a joint optimization of both user association and channel allocation to achieve the best statistical delay performance provisioning.

Although the exact QoS exponent θ value for each user is different, it is known that it satisfies (13), i.e. $R(\theta_{\mathbb{A}_j}, \boldsymbol{n}_{\mathbb{A}_j}) = R(\theta'_{U_i}, \boldsymbol{n}'_{U_i}) = R(\theta_{U_i}, \boldsymbol{n}_{U_i}) = R_{\min}$, hence the objective of the joint user association and channel allocation can be simplified as follows:

$$\max_{j, \mathbb{A}_j} R(\theta_{\min}, \boldsymbol{n}_{\mathbb{A}_j}) + \sum_{U_i \in \mathbb{U}_j} [R(\theta_{\min}, \boldsymbol{n}'_{U_i}) - R(\theta_{\min}, \boldsymbol{n}_{U_i})] \tag{23a}$$

$$\text{s.t.} \quad |\mathbb{A}_j| \leq L_{\max}, \tag{23b}$$

$$R(\theta_{\min}, \boldsymbol{n}_{\mathbb{A}_j}) \geq R_{\min}, \tag{23c}$$

$$R(\theta_{\min}, \boldsymbol{n}'_{U_i}) \geq R_{\min}. \tag{23d}$$

From (23a) – (23d), it can be seen that the calculations in the objective and constraints are unified to the calculation of the aggregated effective rate of each user at the fixed QoS exponent $\theta_{\min}$, whose closed-form expression has been obtained in (16). Note that the effective rate margin in (22a) is defined in a user-wise manner, therefore for scenarios with heterogeneous statistical delay performance requirements, e.g., different delay bound violation probabilities for different users, the above analysis can be directly extended to such cases. Taking advantage of these features, we propose a distributed joint user association and channel allocation algorithm for the F-RAN in the following part.

B. Proposed Distributed Joint User Association and Channel Allocation Algorithm

In this paper, the associations between users and FAPs are determined by the user-centric algorithms instead of the fixed cell-based architecture. For a typical user in the F-RAN, we consider two parameters ρ_{FAP} and ρ_{I}. The parameter ρ_{FAP} defines the maximum distance that FAPs within the range ρ_{FAP} of this user are considered as potential FAPs. Other users near these potential FAPs are all potential interfering users to this typical user, where the parameter ρ_{I} is defined as the maximum distance between a potential FAP and any other user to be considered as a potential interfering user. In general we have $\rho_{\text{FAP}} \leq \rho_{\text{I}}$, which is due to the consideration of the large scale fading effect, where the signal strength attenuates exponentially with distances. With properly assigned ρ_{I}, the weak interfering users will be excluded from (3), which further reduces the computation complexity. To avoid resource exhaustion, one user can connect to a maximum of $L_{\max}$ FAPs.

The serving status information, including served users and their distances, is recorded by a local FAP operating table in

Algorithm 1: Distributed Joint User Association and Channel Allocation Algorithm

1: Initialize the FAP set $\mathbb{A} = \{\text{AP}_1, \ldots, \text{AP}_L\}$ within ρ_{FAP} of the new user from locally cached regional FAP operating table.
2: Initialize the potential channel set $\mathbb{C} = \{1, \ldots, J\}$.
3: Initialize the potential interfering uses set $\mathbb{U} = \{\mathbb{U}_1, \ldots, \mathbb{U}_J\}$, where $\mathbb{U}_j = \{U_1, \ldots, U_{K_j}\}$, $j = 1, \ldots, J$ and U_i is the ith potential interfering user on channel j, $i = 1, \ldots, K_j$.
4: Check existing users' delay performance if the new user is associated on channel j:
  for $j \in \mathbb{C}$ **do**
    for $U_i \in \mathbb{U}_j$ **do**
      Calculate the new effective rate $R(\theta_{\min}, \mathbf{n}'_{U_i})$.
      if $R(\theta_{\min}, \mathbf{n}'_{U_i}) \geq R_{\min}$ **then**
        Calculate the effective rate margin $R^{\text{margin}}_{U_i} = R(\theta_{\min}, \mathbf{n}'_{U_i}) - R(\theta_{\min}, \mathbf{n}_{U_i})$
      else
        Remove this channel from the list $\mathbb{C} = \mathbb{C} \setminus j$.
      end if
    end for
  end for
5: Select the optimal user association and channel allocation according to eq.(23a):
  for $j \in \mathbb{C}$ **do**
    Select $\min\{L, L_{\max}\}$ FAPs from $\mathbb{A}$ to form a potential FAP association set $\mathbb{A}_j$, which achieves the maximal effective rate $R(\theta_{\min}, \boldsymbol{n}_{\mathbb{A}_j})$ on channel j.
    if $R(\theta_{\min}, \boldsymbol{n}_{\mathbb{A}_j}) < R_{\min}$ **then**
      Remove this channel from the list $\mathbb{C} = \mathbb{C} \setminus j$.
    else
      Calculate $R^{\text{overall}}_j = R(\theta_{\min}, \boldsymbol{n}_{\mathbb{A}_j}) + \sum_{U_i \in \mathbb{U}_j} R^{\text{margin}}_{U_i}$.
    end if
  end for
6: Finalize the optimal user association and channel allocation and inform the new user:
  if $\mathbb{C} = \emptyset$ **then**
    The new user is allocated to High Power AP.
  else
    1: The new user is associated with FAP set $\mathbb{A}_j$ on channel j with the largest R^{overall}_j, while the nearest FAP is selected as the main serving FAP.
    2: Update the local and regional FAP operating tables.
  end if

the storage at this FAP. In the meantime, each FAP also maintains a regional FAP operating table, which consists of local FAP operating tables within its range ρ_{regional} via the high-speed connections. This ensures each potential FAP within the range ρ_{FAP} contains all necessary regional information for the joint user association and channel allocation. This information includes all potential interfering users, their distances to these potential FAPs, and their channel in use. The nearest FAP will be selected as the main serving FAP of this new user, who will inform the user of the association result and channel to use.

A distributed joint user association and channel allocation algorithm is proposed in Algorithm 1. The main idea is that each individual FAP exploits the pre-cached FAP operating status to estimate the regional performance based on local computing, while the decision is based on the exhaustive search of all potential user association and channel allocation solutions. It can be seen that during the decision procedure, each potential FAP carries out the same calculation with the cached regional FAP information. This enables the distributed and local decision making, which saves the decision procedure from iterative inquiry between the FAPs or centralized calculations. As the user number K is the key parameter defining the total solution space, the algorithm's time complexity is further reduced to $\mathcal{O}(K)$, i.e. linear complexity, which will be detailed in the following subsection.

C. Computing Complexity and Storage Considerations

A key feature of F-RAN is the distributed yet limited computing resources and storage capacity at each FAPs. In this part, the computing complexity and storage requirement of the proposed joint user association and channel allocation algorithm will be discussed. As the effective rate function is the main metric in Algorithm 1, the following efforts are made to reduce its computing complexity and storage requirement, which is achieved by the transformation of multivariate *H* function calculation to the look-up table operation.

We start from the revisit of (16), which describes the effective rate of U_0 with multiple FAP associations and arbitrary interfering users in a closed form. As seen from its expression, (16) can be decoupled to the calculation of individual multi-variate *H* functions. Each multi-variate *H* function is well sorted according to the subscript index $k_l, l = 1, \ldots, L$, which corresponds to the interfering user U_{k_l}. In this way, the calculation of the aggregated effective rate can be decoupled to the calculation of multi-variate *H* function based on each involved interfering users.

In general, the multi-variate *H* function can be evaluated using Python [31], Matlab with C/MEX or GPU accelerations [32]. As this calculation is frequently involved in the estimation of effective rate function, in the following, we will provide a loop-up table based solution with a low complexity and storage requirement. There are four key observations as follows: Firstly, for a given user, there are L variables in the multi-variate *H* function in (16), which corresponds to its associated FAP number. Secondly, when the scenarios is changed for an L-variate *H* function in (16), it only changes the variable's value $\frac{n_{l0}}{n_{lk_l}}$ and the ratio factor w_{lk_l}, while the other parameters are constant. Thirdly, the variable order can be swapped due to the multi-variate *H* function definition in Appendix A. Finally, each multi-variate *H* function in (16) is monotonically decreasing with respect to each variable according to Lemma 1. Using these features, we can sample each L-variate *H* function into a table with $w_{lk_l} = 1$. Given a calculation request with n_{l0} and $[n_{lk_1}, \ldots, n_{lk_L}]$, we can use linear interpolation to obtain the L-variate *H* function at $[\frac{n_{l0}}{n_{lk_1}}, \ldots, \frac{n_{l0}}{n_{lk_L}}]$ via look-up table, and then multiply the ratio factor $\prod_{l=1}^{L} w_{lk_l}$. Since the interpolation and look-up tables are independent of the user number K, the time complexity of multi-variate *H* function can be estimated by $\mathcal{O}(1)$.

Considering the storage limit, the resolution of the sampling procedure can be adjusted according to the accuracy and local storage capacity. As illustrated in Fig. 2, by exponentially sampling the multi-variate *H* functions, the interpolated value asymptotically approaches the exact values with the increase

of sampling resolution. For a sampling resolution of $10^{0.5}$, it can already achieve a good approximation performance.

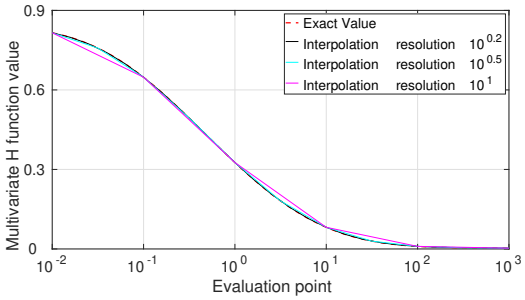

Fig. 2. The interpolation and look-up table performance of the multi-variate H function with two variables.

When adding a new interfering user, (16) shows that the new effective rate function will only add a maximum of K new summation items, while the previously calculated multi-variate H functions can be re-used with an adjusted w_{lk_l}. Therefore the estimated time complexity in solving (16) is reduced from $\mathcal{O}(K!)$ (due to the enumeration of all possible combinations of $k_1, \ldots, k_L$) to $\mathcal{O}(K)$ by reusing the previous results. As illustrated in Fig. 3, under the same FAP density, the time performance shows a linear time complexity with regard to the user density.

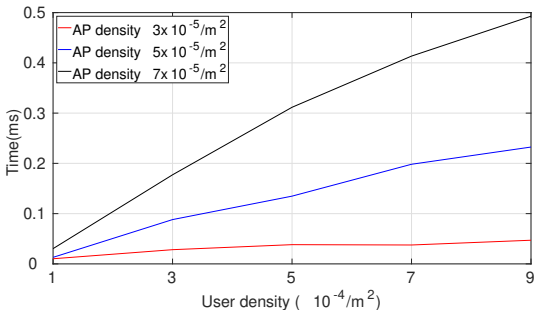

Fig. 3. The time performance of the algorithm under different user densities.

Therefore from the aspect of both computing complexity and storage capacity, the proposed joint user association and channel allocation algorithm and scheme suggests promising potentials to be fulfilled by the local computation at the distributed FAPs. In this way, the computation-communications tradeoff is addressed via the distributed Algorithm 1 with low complexity considerations and the regional operating information updates via regional FAP communications.

D. Network Association Protocol

The association between users and FAPs involves high control signaling overhead, which has a great impact on the overall delay performance. The time step diagram of the proposed distributed association scheme is given on the right of Fig. 4. The user sends the network association request in the control channels, which contains the position information such as Global Positioning System (GPS) data. The FAPs in the surroundings listen to this control channel and then execute

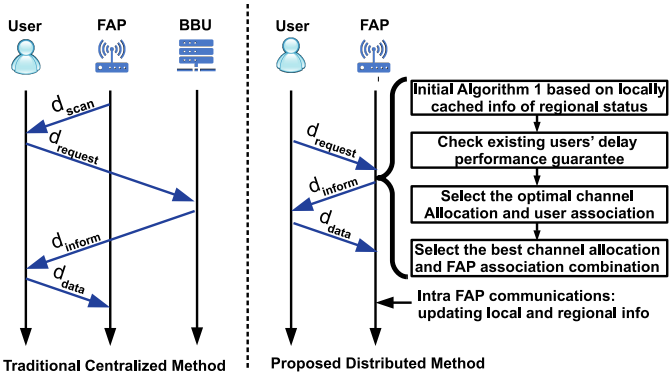

Fig. 4. The time-step diagram of the proposed distributed method, comparing to the traditional centralized method.

the Algorithm 1 to make the decision on FAP associations and channel allocations. The selected main serving FAP informs the user about the results, while at this time point the regional FAP cluster is already formed via the local computation. After that, the user commences the data transmission on the informed channel. Comparing to the traditional centralized association method as illustrated on the left of Fig. 4, it can be seen that the proposed method has simplified the association procedure with fewer control signaling steps, which leads to a reduction of the total delay overheads. There are two key changes compared to traditional closed-loop association schemes. Firstly, the decision is made on the FAPs locally, which saves the time to consult the BBUs and wait for the response. Secondly, although the FAP cluster is formed in a user-centric way, the users do not need to care about the available FAPs. The proposed Algorithm 1 dynamically adjusts the associated FAPs to the users according to the interfering conditions. If an existing user has changed its position outside ρ_{I} of an FAP, then it can exploit Algorithm 1 to update the local and regional operating status.

E. Delay-Guarantee Necessity Condition

The joint user association and channel allocation scheme is designed to protect the delay performance of both new and existing users, while the latency for individual users is determined by the general communication resources, including FAP density λ_{FAP}, user density λ_{User} and total available channels J. It is intuitive that there exists a necessary condition under which the delay can be guaranteed. From an averaged aspect, the involved parameter can be bounded by the following necessary bound to serve the users with only FAP connections.

Proposition 3: For a F-RAN network with an average of $L^{\dagger}$ available FAPs and $K^{\dagger}$ served users per channel, a necessary bound for the statistical delay-guaranteed of the proposed joint user association and channel allocation algorithm can be given as follows:

$$-\frac{1}{A^{\dagger}}\log_2\frac{1}{\Gamma(A^{\dagger})}H_{1,0}^{0,1}\left[\begin{matrix}(1-A^{\dagger},\mathbf{1}_{L^{\dagger}}):\\ —\end{matrix}\left(\frac{1}{L^{\dagger}},\mathbf{O}^{\dagger},\mathbf{P}^{\dagger}\right)_{1,L^{\dagger}}\right] \geq R_{\min}, \tag{24}$$

where $A^\dagger = \frac{\theta_{\min} TB}{\ln 2}$, $\mathbf{O}^\dagger = [2,1,1,2]$, $\mathbf{P}^\dagger = [\frac{K^\dagger}{\Gamma(K^\dagger)}, \left(\frac{r_{\text{avg}}}{\rho_{\text{I}}}\right)^{-\alpha}, 0, (0, K^\dagger - 1), 1, (1,1)]$ and r_{avg} is the average distance between FAPs and users.

Proof: See Appendix C. ∎

The equation (24) can be used to estimate the upper bound of the FAP density λ_{FAP} and the user density λ_{User} under specific scenarios. If the FAPs and users are uniformly distributed, then the average number of available FAPs is estimated as $L^\dagger \approx \text{floor}(\pi \rho_{\text{FAP}}^2 \lambda_{\text{FAP}})$, the average number of served user per channel by each FAP is estimated as $K^\dagger \approx \text{ceil}(\frac{\pi \rho_{\text{I}}^2 \lambda_{\text{User}}}{J})$, while the average distance between FAP and users is estimated as $r_{\text{avg}} \approx \frac{2\rho_{\text{FAP}}}{3}$. For other FAP and user distributions, (24) can be also used with corresponding parameters.

Note that the metrics in Algorithm 1 are based on the worst-case analysis, i.e., assuming that all interfering users are simultaneously transmitting signals. This will ensure that the delay-guarantee will be applicable to all practical scenarios, even for the worst case. For example, when there are only two users U_1 and U_2 on the same channel and associated with the same cluster of FAPs, the delay performance of both U_1 and U_2 is guaranteed assuming they are transmitting at the same time. For the case U_1 is transmitting while U_2 is not, it is expected that U_1 will have a better statistical delay performance than the guaranteed metric.

IV. Simulation Results

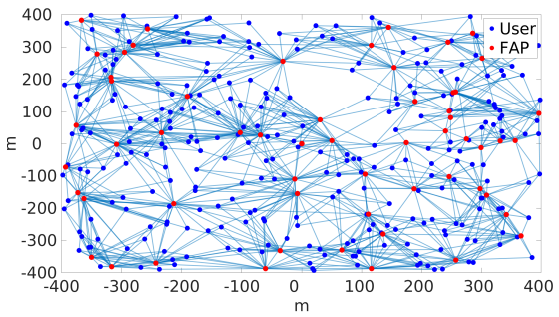

Fig. 5. An example of the user associations with FAP density of $9\times10^{-5}/\text{m}^2$ and user density of $3 \times 10^{-4}/\text{m}^2$.

In this section, the proposed joint user association and channel allocation scheme will be evaluated by simulations. It is assumed that the frame duration is $T = 1$ ms, the bandwidth for each channel is $B = 10$ kHz and the average package arrival rate is $\lambda = 50$ packets/s. The maximum delay bound is $D_{\max} = 8$ ms, while the SIR threshold is $\gamma = 3$ to support the QPSK transmission [16]. The path-loss exponent is set as $\alpha = 6$ to account for the obstruction effect of the buildings. Correspondingly, the association range is set as $\rho_{\text{FAP}} = 200$ m and the interfering range is set as $\rho_{\text{I}} = 240$ m. The maximum association number $L_{\max}$ is set as 4. The simulation scenario is illustrated in Fig. 5, where the association between users and FAPs is also presented. Two benchmark algorithms are used, which are the Horizontal Network Association algorithm in [16] with the control parameter $V = 0.01$ due to its best-reported performance in the literature, and the Multi-Connectivity algorithm in [33] with $L_{\max} = 4$, whose computation complexities are both $\mathcal{O}(1)$. Different user densities and FAPs densities are simulated, whose distributions follow the Poisson Point Process (PPP).

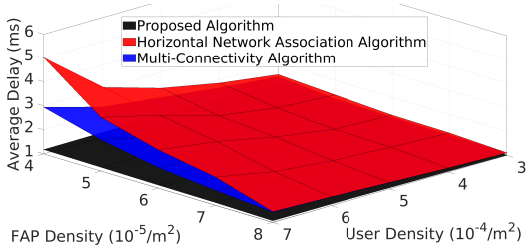

Fig. 6. The average delay performance under different FAP and user densities.

In Fig. 6, the average delay performance is evaluated against both FAP density and user density. It can be seen that the proposed joint user association and channel allocation algorithm outperforms the benchmark algorithms and achieves lower average delay across the considered scenarios. For relatively low user density (e.g., the range of $3 \times 10^{-4}/\text{m}^2$ to $5 \times 10^{-4}/\text{m}^2$) and high FAP density (e.g., the range of $6 \times 10^{-5}/\text{m}^2$ to $8 \times 10^{-5}/\text{m}^2$), all algorithms can achieve an average delays within 2ms. With more users and less FAPs, the proposed algorithm is less sensitive to the changes and the average delay is maintained under 2ms, comparing to the benchmark algorithms.

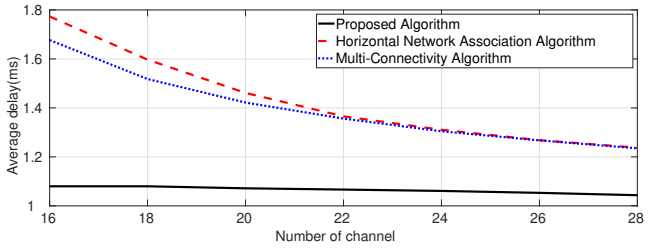

Fig. 7. The average delay performance under different total number of channels, where the user density and FAP density are $5 \times 10^{-4}/\text{m}^2$ and $7 \times 10^{-5}/\text{m}^2$, respectively.

The average delay performance under different number of channels is presented in Fig. 7, where the user density and FAP density are $5 \times 10^{-4}/\text{m}^2$ and $7 \times 10^{-5}/\text{m}^2$, respectively. During the joint user association and channel allocation, the interference between users can be spread among all channels via the proposed algorithm. Therefore the proposed algorithm shows a better performance against the benchmark algorithms whose channels are randomly allocated. Especially when the number of channel is reduced, chances increase when the users are allocated to channels with more interference, while it is mitigated in the proposed algorithm as the interference on each channel is estimated.

With the same FAP density of $9\times10^{-5}/\text{m}^2$, the performance of the average connected FAP for each user is presented in red lines and crossing marks in Fig. 8, while the performance of the average served users by an individual FAP on each channel is illustrated in blue lines and round marks. It can be observed that the proposed algorithm is efficient regarding the usage of channel and FAP resources, which corresponds to the

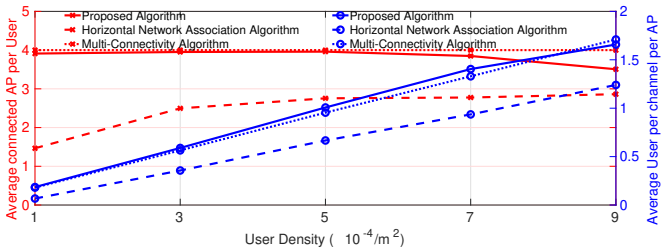

Fig. 8. The network utility performance: a) the average associated FAP number in serving each user, and b) the average served user served by individual FAP on each channel.

most served user number per channel per AP below the user density $7 \times 10^{-4}/\mathrm{m}^2$. For user density above $9 \times 10^{-4}/\mathrm{m}^2$, this metric dropped below the Multi-Connectivity algorithm. This is because in the proposed algorithm, the performance of both existing users and the new user is considered. The number of served users per channel per FAP is reduced to mitigate the interference increase, which is different from the benchmark algorithms showing linear dependency to the user density. This is also observed via the average connected AP per user performance, which decreases as the user density grows.

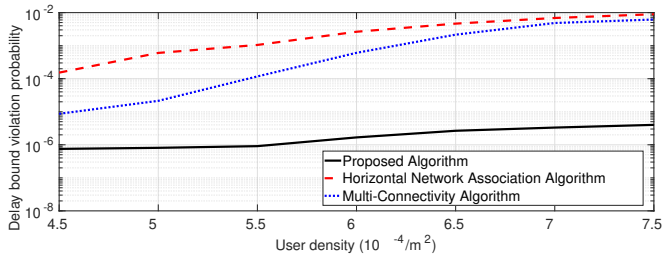

Fig. 9. The delay bound violation probability across different FAP densities.

The delay bound violation performance under FAP density $3 \times 10^{-5}/\mathrm{m}^2$ is presented in Fig. 9. Comparing to the benchmark algorithms, the proposed joint user association and channel allocation algorithm shows better statistical delay bound guarantee performance, which is below 1×10^{-5} across all simulated scenarios. This is because in the benchmark algorithm, the channel is randomly allocated to the users. There are cases in which multiple users are allocated to the same channel and experience high collisions, while the other channels are not effectively used. In the proposed algorithm, this phenomenon is mitigated since the delay bound violation probability is estimated for the new user with the consideration of the existing statuses of both users and FAPs, while the new user's impact on existing users' delay performance is also estimated and protected during the user association and channel allocation procedure. From the view of system design for 5G and beyond networks, the regional coordination among distributed FAPs shows a promising solution to meet the ultra-low latency requirement. The distributed algorithm also demonstrates that the FAP's local computing and storage capacity have potentials in further pushing the communication services to the network edge, which would benefit other future network designs such as edge-caching and edge-computing.

V. Conclusions

In this paper, the statistical delay performance of the users under MAI in the F-RAN scenario has been studied. The effective rate of a typical user with multiple FAP connections and arbitrary interfering users was obtained in a closed form. Based on these analyses, a novel joint user association and channel allocation algorithm was proposed, which considered not only existing users' impact on the achievable performance of the new user, but also the new user's influence on the existing users. By taking advantage of F-RAN architecture and the local computing and storage resources on the FAPs, the algorithm was decoupled to facilitate distributed computing. The computing complexity and storage constraints for distributed FAPs were also addressed by look-up tables operations. The proposed scheme improved the delay performance with regard to both control signaling procedure and the data transmission procedure. Simulation results demonstrated that the proposed joint user association and channel allocation scheme was promising in addressing statistical delay provisioning problems in the F-RAN scenarios. In the future work, we will extend the scheme to more general fading scenarios and the distributed MIMO systems with multi-antenna users.

Appendices

A. Multi-variate Fox's H function

The multi-variate H function [19, Appendix A.1] is defined by multiple Mellin-Barnes type contour integrals as follows:

$$\begin{aligned} & H_{p_0,q_0}^{0,n_0}\left[\begin{matrix}(a_j;A_j^{(1)},\ldots,A_j^{(N)})_{p_0}:(s_j,\mathbf{O}_j,\mathbf{P}_j)_N\\(b_j;B_j^{(1)},\ldots,B_j^{(N)})_{q_0}\end{matrix}\right] \\ =&\frac{u_1\cdots u_N}{(2\pi\sqrt{-1})^N}\int_{\mathcal{L}_1}\cdots\int_{\mathcal{L}_N}\Psi(\zeta_1,\ldots,\zeta_N)\times \\ &\left\{\prod_{j=1}^{N}\Phi_j(\zeta_j)(v_j s_j)^{\zeta_j}\right\}d\zeta_1\cdots d\zeta_N, \end{aligned} \tag{25}$$

where $\mathcal{L}_j$ is the suitable contours in the ζ_j-plane. $(s_j,\mathbf{O}_j,\mathbf{P}_j)_N$ abbreviates N-parameter array $(s_1,\mathbf{O}_1,\mathbf{P}_1);\ldots;(s_N,\mathbf{O}_N,\mathbf{P}_N)$. $\mathbf{O}_j = (m_j,n_j,p_j,q_j)$ and $\mathbf{P}_j = (u_j,v_j,\mathbf{c}^{(j)},\mathbf{d}^{(j)},\mathbf{C}^{(j)},\mathbf{D}^{(j)})$, whereas $\mathbf{c}^{(j)}$ abbreviates p_j-parameter array $(c_1^{(j)},\ldots,c_{p_j}^{(j)})$, and

$$\Psi(\zeta_1,\ldots,\zeta_N) = \frac{\prod_{j=1}^{n_0}\Gamma(1-a_j+\sum_{\ell=1}^{N}A_j^{(\ell)}\zeta_\ell)}{\prod_{j=n_0+1}^{p_0}\Gamma(a_j-\sum_{\ell=1}^{N}A_j^{(\ell)}\zeta_\ell)\prod_{j=1}^{q_0}\Gamma(1-b_j+\sum_{\ell=1}^{N}B_j^{(\ell)}\zeta_\ell)}, \tag{26}$$

$$\Phi_j(\zeta_j)=\frac{\prod_{\ell=1}^{m_j}\Gamma(d_\ell^{(j)}-D_\ell^{(j)}\zeta_j)\prod_{\ell=1}^{n_j}\Gamma(1-c_\ell^{(j)}+C_\ell^{(j)}\zeta_j)}{\prod_{\ell=n_j+1}^{p_j}\Gamma(c_\ell^{(j)}-C_\ell^{(j)}\zeta_j)\prod_{\ell=m_j+1}^{q_j}\Gamma(1-d_\ell^{(j)}+D_\ell^{(j)}\zeta_j)}, \tag{27}$$

For $N = 1$, the multi-variate H function reduces to the univariate H function [19, Ch. 1.2] as follows:

$$uH_{p,q}^{m,n}\left[vs\middle|\begin{matrix}\mathbf{c},\mathbf{C}\\\mathbf{d},\mathbf{D}\end{matrix}\right]=\frac{u}{2\pi\sqrt{-1}}\int_{\mathcal{L}}\Phi(\zeta)(vs)^{\zeta}d\zeta, \tag{28}$$

where the parameters are following similar definitions in multi-variate H function with subscripts omitted [28]. In this paper the unified representation of $H_{p,q}^{m,n}[\cdot]$ is used for both multi-variate and uni-variate H function, which can be distinguished by the number of involved random variables.

B. Proof of Proposition 1

Using $f_{I_{lk_l}}(z)$ in (6), the PDF of the aggregated I_l can be given as follows with the consideration of their mutual independency [29]:

$$f_{I_l}(z) = \sum_{k_l=1}^{K_l} \frac{w_{lk_l}}{n_{lk_l}} e^{-\frac{z}{n_{lk_l}}}. \tag{29}$$

Using the joint distribution theorem in [34, Theorem 2-6] and the identity $\int_0^\infty x^n e^{-\mu x} dx = n\mu^{-n-1}$ [35, (3.351.3)], the PDF of γ_l can be given by

$$f_{\gamma_l}(x) = \sum_{k_l=1}^{K_l} w_{lk_l} \frac{n_{l0}}{n_{lk_l}} (x + \frac{n_{l0}}{n_{lk_l}})^{-2}. \tag{30}$$

By substituting (30) to the MGF definition $\phi_{\gamma_l}(s) = \mathbb{E}_{\gamma_l}\{e^{-s\gamma_l}\}$, and using a) the integral definition of Kummer U function [19, (13.4.4)], b) its relation with generalized hypergeometric functions [19, (13.6.21)] and [19, (16.18.1)], and c) the definition of Meijer-G function definition [19, (1.111)], the closed-form MGF in (7) can be obtained.

C. Proof of Proposition 3

For a given user, its distance to a FAP is r_{avg} , while there are $K^\dagger - 1$ co-channel users are to be served within an FAP's range ρ_{FAP}. Then the best condition this user can achieve at this FAP is that all the other $K^\dagger - 1$ users are located at the furthest position from the FAP introducing the least interference, i.e., all with the distance of ρ_{I}. In this case the PDF of the aggregated interference on the FAP due to the $K^\dagger - 1$ users can be given by

$$f_I(x) = K^\dagger (\frac{r_{\text{avg}}}{\rho_{\text{I}}})^\alpha \left(1 + x(\frac{r_{\text{avg}}}{\rho_{\text{I}}})^\alpha\right)^{-K^\dagger}. \tag{31}$$

Furthermore, this user will achieve its best aggregated SIR performance and hence the aggregated effective rate performance by connecting all $L^\dagger$ available FAPs within its range. If all $L^\dagger$ FAPs are in the similarly best condition and the user still cannot reach the stable condition of $R(\theta^\dagger, \boldsymbol{n}^\dagger) \geq R_{\min}$, then the network cannot serve the users with only FAPs from an average aspect. With (31) and the same procedure as Proposition 1 and Proposition 2, the left hand side of (24) can be achieved. Then by applying Lemma 1, (24) is obtained.

REFERENCES

[1] S. Andreev, V. Petrov, M. Dohler, and H. Yanikomeroglu, "Future of ultra-dense networks beyond 5G: Harnessing heterogeneous moving cells," *IEEE Commun. Mag.*, vol. 57, no. 6, pp. 86–92, Jun. 2019.

[2] C. Zeng, K. Chen, and D. Liu, "Two-stage ICI suppression in the downlink of asynchronous URLLC," *IEEE Trans. Wireless Commun.*, vol. 19, no. 4, pp. 2785–2799, Apr. 2020.

[3] G. M. S. Rahman, M. Peng, K. Zhang, and S. Chen, "Radio resource allocation for achieving ultra-low latency in fog radio access networks," *IEEE Access*, vol. 6, pp. 17 442–17 454, Feb. 2018.

[4] M. Peng, Y. Sun, X. Li, Z. Mao, and C. Wang, "Recent advances in cloud radio access networks: System architectures, key techniques, and open issues," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 3, pp. 2282–2308, Mar. 2016.

[5] M. Peng, Y. Li, J. Jiang, J. Li, and C. Wang, "Heterogeneous cloud radio access networks: a new perspective for enhancing spectral and energy efficiencies," *IEEE Wireless Commun.*, vol. 21, no. 6, pp. 126–135, Dec. 2014.

[6] T. Chiu, A. Pang, W. Chung, and J. Zhang, "Latency-driven fog cooperation approach in fog radio access networks," *IEEE Trans. Services Comput.*, pp. 1–14, Jul. 2018.

[7] Cisco Global Cloud Index: Forecast and Methodology, 2016C2021 White Paper. (Accessed: Nov. 1, 2020). [Online]. Available: https://www.cisco.com/c/en/us/solutions/collateral/service-provider/global-cloud-indexgci/white-paper-c11-738085.html

[8] R. Deng, R. Lu, C. Lai, T. H. Luan, and H. Liang, "Optimal workload allocation in fog-cloud computing toward balanced delay and power consumption," *IEEE Internet Things J.*, vol. 3, no. 6, pp. 1171–1181, Dec. 2016.

[9] X. Ge, X. Li, H. Jin, J. Cheng, and V. C. Leung, "Joint user association and user scheduling for load balancing in heterogeneous networks," *IEEE Trans. Wireless Commun.*, vol. 17, no. 5, pp. 3211–3225, Feb. 2018.

[10] S. Bassoy, M. Jaber, M. A. Imran, and P. Xiao, "Load aware self-organising user-centric dynamic CoMP clustering for 5G networks," *IEEE Access*, vol. 4, pp. 2895–2906, May 2016.

[11] W. Jing, X. Wen, Z. Lu, and H. Zhang, "User-centric delay-aware joint caching and user association optimization in cache-enabled wireless networks," *IEEE Access*, vol. 7, pp. 74 961–74 972, May 2019.

[12] X. Luo, "Delay-oriented QoS-aware user association and resource allocation in heterogeneous cellular networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1809–1822, Jan. 2017.

[13] W. Wang, V. K. N. Lau, and M. Peng, "Delay-aware uplink fronthaul allocation in cloud radio access networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 7, pp. 4275–4287, Jul. 2017.

[14] Z. Li, J. Chen, L. Zhen, S. Cui, K. G. Shin, and J. Liu, "Coordinated multi-point transmissions based on interference alignment and neutralization," *IEEE Trans. Wireless Commun.*, vol. 18, no. 7, pp. 3347–3365, Jul. 2019.

[15] Y. Sun, Z. Ding, X. Dai, and O. A. Dobre, "On the performance of network NOMA in uplink CoMP systems: A stochastic geometry approach," *IEEE Trans. Commun.*, vol. 67, no. 7, pp. 5084–5098, Jul. 2019.

[16] S. Hung, H. Hsu, S. Cheng, Q. Cui, and K. Chen, "Delay guaranteed network association for mobile machines in heterogeneous cloud radio access network," *IEEE Trans. Mobile Comput.*, vol. 17, no. 12, pp. 2744–2760, Mar. 2018.

[17] S.-C. Hung, H. Hsu, S.-Y. Lien, and K.-C. Chen, "Architecture harmonization between cloud radio access networks and fog networks," *IEEE Access*, vol. 3, pp. 3019–3034, Dec. 2015.

[18] F. W. Olver, D. W. Lozier, R. F. Boisvert, and C. W. Clark, *NIST handbook of mathematical functions*. New York, NY, USA: Cambridge University Press, 2010.

[19] A. M. Mathai, R. K. Saxena, and H. J. Haubold, *The H-function: theory and applications*, 2010th ed. New York, USA: Springer Science & Business Media, 2009.

[20] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, "Massive MIMO for next generation wireless systems," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 186–195, Feb. 2014.

[21] J. Tang and X. Zhang, "Cross-layer modeling for quality of service guarantees over wireless links," *IEEE Trans. Wireless Commun.*, vol. 6, no. 12, pp. 4504–4512, Dec. 2007.

[22] D. Qiao, M. Gursoy, and S. Velipasalar, "Effective capacity of two-hop wireless communication systems," *IEEE Trans. Inf. Theory*, vol. 59, no. 2, pp. 873–885, Feb. 2013.

[23] M. You, J. Jiang, A. M. Tonello, T. Doukoglou, and H. Sun, "On statistical power grid observability under communication constraints," *IET Smart Grid*, vol. 1, no. 2, pp. 40–47, Aug. 2018.

[24] M. Amjad, L. Musavian, and M. H. Rehmani, "Effective capacity in wireless networks: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 4, pp. 3007–3038, Jul. 2019.

[25] D. Wu and R. Negi, "Effective capacity: a wireless link model for support of quality of service," *IEEE Trans. Wireless Commun.*, vol. 2, no. 4, pp. 630–643, Jul. 2003.

[26] M. You, H. Sun, J. Jiang, and J. Zhang, "Effective rate analysis in weibull fading channels," *IEEE Wireless Commun. Lett.*, vol. 5, no. 4, pp. 340–343, Aug. 2016.
[27] J. Choi, "An effective capacity-based approach to multi-channel low-latency wireless communications," *IEEE Trans. Wireless Commun.*, vol. 67, no. 3, pp. 2476–2486, Mar. 2019.
[28] M. You, H. Sun, J. Jiang, and J. Zhang, "Unified framework for the effective rate analysis of wireless communication systems over MISO fading channels," *IEEE Trans. Commun.*, vol. 65, no. 4, pp. 1775–1785, Apr. 2017.
[29] I. S. Ansari, F. Yilmaz, M.-S. Alouini, and O. Kucur, "New results on the sum of Gamma random variates with application to the performance of wireless communication systems over Nakagami-*m* fading channels," *Trans. Emerg. Telecommun. Technol.*, vol. 28, no. 1, pp. 1–14, Dec. 2017.
[30] IEEE Vehicular Technology Society Committee on Radio Propagation, "Coverage prediction for mobile radio systems operating in the 800/900 MHz frequency range," *IEEE Trans. Veh. Technol.*, vol. 37, no. 1, pp. 3–72, Feb. 1988.
[31] H. R. Alhennawi, M. M. H. El Ayadi, M. H. Ismail, and H. M. Mourad, "Closed-form exact and asymptotic expressions for the symbol error rate and capacity of the *H*-function fading channel," *IEEE Trans. Veh. Technol.*, vol. 65, no. 4, pp. 1957–1974, Apr. 2016.
[32] H. Chergui, M. Benjillali, and M. Alouini, "Rician *K* factor based analysis of XLOS service probability in 5G outdoor ultra-dense networks," *IEEE Wireless Commun. Lett.*, vol. 8, no. 2, pp. 428–431, Apr. 2019.
[33] H. Zhang, W. Huang, and Y. Liu, "Handover probability analysis of anchor-based multi-connectivity in 5G user-centric network," *IEEE Wireless Commun. Lett.*, vol. 8, no. 2, pp. 396–399, Apr. 2019.
[34] J. S. Milton and J. C. Arnold, *Schaum's Outline of Introduction to Probability & Statistics: Principles & Applications for Engineering & the Computing Sciences*. New York, USA: McGraw-Hill Higher Education, 1994.
[35] I. S. Gradshteyn and I. M. Ryzhik, *Table of integrals, series, and products*, 7th ed. Amsterdam: Elsevier/Academic Press, 2007.

Minglei You (S'15) received his PhD degree from the University of Durham (U.K.) in 2019 and master degree from the Beijing University of Posts and Telecommunications (China) in 2014. In 2012, he was a short-term visiting student at the University of Electro-Communications (Japan). Since 2014, he has been with the University of Durham as a recipient of the Durham Doctoral Scholarship. From 2018 to 2019, he was a Postdoctoral Research Associate with the Loughborough University (U.K.). He is currently a Postdoctoral Research Associate with the University of Durham. His recent research interest includes machine learning for communications, testbed design, Smart Grid and cyber security.

Gan Zheng (S'05–M'09–SM'12–F'21) received the BEng and the MEng from Tianjin University, Tianjin, China, in 2002 and 2004, respectively, both in Electronic and Information Engineering, and the PhD degree in Electrical and Electronic Engineering from The University of Hong Kong in 2008. He is currently Reader of Signal Processing for Wireless Communications in the Wolfson School of Mechanical, Electrical and Manufacturing Engineering, Loughborough University, UK. His research interests include machine learning for communications, wireless power transfer, UAV communications, mobile edge caching and full-duplex radio. He is the first recipient for the 2013 IEEE Signal Processing Letters Best Paper Award, and he also received 2015 GLOBECOM Best Paper Award, and 2018 IEEE Technical Committee on Green Communications & Computing Best Paper Award. He was listed as a Highly Cited Researcher by Thomson Reuters/Clarivate Analytics in 2019. He currently serves as an Associate Editor for IEEE Communications Letters and IEEE Wireless Communications Letters. He is Fellow of IEEE.

Tianrui Chen received the MSc degree in communication engineering from Durham University, Durham, U.K., in 2018. She is currently pursuing the Ph.D. degree with the Signal Processing and Networks Research Group, Wolfson School of Mechanical, Electrical and Manufacturing Engineering, Loughborough University, U.K. Her research interests include resource allocation in 5G and beyond networks, D2D networks, blockchain technologies and machine learning for communications.

Hongjian Sun (S'07–M'11–SM'15) received his Ph.D. degree at the University of Edinburgh (U.K.) in 2011 and then took postdoctoral positions at Kings College London (U.K.) and Princeton University (USA). Since April 2013, he has been with the Department of Engineering at the University of Durham (U.K.) as Professor (Chair) since July 2020, Associate Professor (Reader) in 2017-2020, and Assistant Professor in 2013-2017. He has published over 120 papers in refereed journals and international conferences; He has made contributions to and coauthored the IEEE 1900.6a-2014 Standard; In addition, he has published 5 book chapters, and edited 2 books. He is the Editor-in-Chief of the IET Smart Grid Journal and is Principle Investigator or Co-Investigator on a number of projects, with funding successes from the EU H2020, EU ERDF, UK EPSRC, UK BEIS, Innovate UK, China NSF, and industry. He is a Fellow of Durham Energy Institute, and a Fellow of Higher Education Academy.

Kwang-Cheng Chen (F'07) received the B.S. from the National Taiwan University in 1983, and the M.S. and Ph.D from the University of Maryland, College Park, United States, in 1987 and 1989, all in electrical engineering. From 1987 to 1998, Dr. Chen worked with SSE, COMSAT, IBM Thomas J. Watson Research Center, and National Tsing Hua University. From 1998 to 2016, he was a Distinguished Professor with the National Taiwan University, Taipei, Taiwan, ROC, also served as the Director, Graduate Institute of Communication Engineering, Director, Communication Research Center, and the Associate Dean for Academic Affairs, College of Electrical Engineering and Computer Science during 2009-2015. Since 2016, Dr. Chen has been the Professor of Electrical Engineering, University of South Florida, Tampa, Florida. He has been actively involving in the organization of various IEEE conferences as General/TPC chair/co-chair, and has served in editorships with a few IEEE journals. Dr. Chen also actively participates in and has contributed essential technology to various IEEE 802, Bluetooth, LTE and LTE-A, 5G-NR, and ITU-T FG ML5G wireless standards. Dr. Chen is an IEEE Fellow and has received a number of awards including the *2011 IEEE COMSOC WTC Recognition Award*, *2014 IEEE Jack Neubauer Memorial Award* and *2014 IEEE COMSOC AP Outstanding Paper Award*. His recent research interests include wireless networks, multi-robot systems, IoT and CPS, social networks and data analytics, and cybersecurity.